Discovery of potential causative mutations in human coding and noncoding genome with the interactive software BasePlayer

Katainen, Riku

2018-11

Katainen , R , Donner , I , Cajuso , T , Kaasinen , E , Palin , K , Mäkinen , V , Aaltonen , L A & Pitkänen , E 2018 , ' Discovery of potential causative mutations in human coding and noncoding genome with the interactive software BasePlayer ' , Nature Protocols , vol. 13 , no. 11 , pp. 2580-2600 . https://doi.org/10.1038/s41596-018-0052-3

http://hdl.handle.net/10138/306620
https://doi.org/10.1038/s41596-018-0052-3

publishedVersion

nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-018-0052-3

Discovery of potential causative mutations in human coding and noncoding genome with the interactive software BasePlayer

Riku Katainen[1,2*], Iikki Donner[1,2], Tatiana Cajuso[1,2], Eevi Kaasinen[1,2], Kimmo Palin[1,2], Veli Mäkinen[3], Lauri A. Aaltonen[1,2], Esa Pitkänen[1,2,4*]

Next-generation sequencing (NGS) is routinely applied in life sciences and clinical practice, but interpretation of the massive quantities of genomic data produced has become a critical challenge. The genome-wide mutation analyses enabled by NGS have had a revolutionary impact in revealing the predisposing and driving DNA alterations behind a multitude of disorders. The workflow to identify causative mutations from NGS data, for example in cancer and rare diseases, commonly involves phases such as quality filtering, case–control comparison, genome annotation, and visual validation, which require multiple processing steps and usage of various tools and scripts. To this end, we have introduced an interactive and user-friendly multi-platform-compatible software, BasePlayer, which allows scientists, regardless of bioinformatics training, to carry out variant analysis in disease genetics settings. A genome-wide scan of regulatory regions for mutation clusters can be carried out with a desktop computer in ~10 min with a dataset of 3 million somatic variants in 200 whole-genome-sequenced (WGS) cancers.

Introduction

DNA alterations have been studied for decades to unravel the mechanisms underlying disease development as well as cellular and evolutionary processes. NGS technologies have increased the scale of genetic studies by multiple orders of magnitude and are at present widely applied in the life sciences. These developments have resulted in the availability of massive amounts of sequencing data for research and clinical use[1]. A fundamental challenge in disease genetics is the identification of a variant or variants causative of the disease or phenotype. Putative causative mutations in the germ line typically possess characteristics such as low allele frequency in the population, overrepresentation in cases versus controls, and, in the case of coding mutations, a damaging effect on the relevant protein[2]. Although the genetic basis of many of the ~7,000 rare diseases affecting >300 million people worldwide is still unknown, NGS is rapidly accelerating the discovery of the missing determinants[2]. At the same time, NGS is driving cancer genetics and genomics. For instance, one can utilize evidence of selection, such as clonality and unexpectedly high degrees of mutation recurrence in a gene or genomic region, to detect somatic driver mutations in cancer[3]. NGS techniques are also rapidly highlighting the role of somatic mutations in other disorders, such as clonal hematopoiesis[4].

Owing to numerous error-prone experimental and computational steps that must be taken to obtain variants from DNA or RNA sequences[5], variant calling is often imperfect[6]. Variant calling is particularly difficult in low-complexity and homologous genomic regions, which are abundant in the noncoding genome, thus complicating the analysis of noncoding variants[7]. In addition, variant calling can be hampered by sample heterogeneity, subclonality and copy-number variation, among other factors[8]. Well-established workflows, such as Broad GATK, facilitate primary analysis of sequence data by providing variant call format (VCF) files annotated with quality information[9] (Fig. 1). NGS data providers such as core facilities and companies commonly provide analysis-ready sequence alignment (BAM) and VCF files. However, filtering by quality values alone is often insufficient, and

[1]Genome-Scale Biology Research Program, Research Programs Unit, University of Helsinki, Helsinki, Finland. [2]Department of Medical and Clinical Genetics, Medicum, University of Helsinki, Helsinki, Finland. [3]Department of Computer Science and Helsinki Institute for Information Technology, University of Helsinki, Helsinki, Finland. [4]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. *e-mail: riku.katainen@helsinki.fi; esa.pitkanen@cs.helsinki.fi

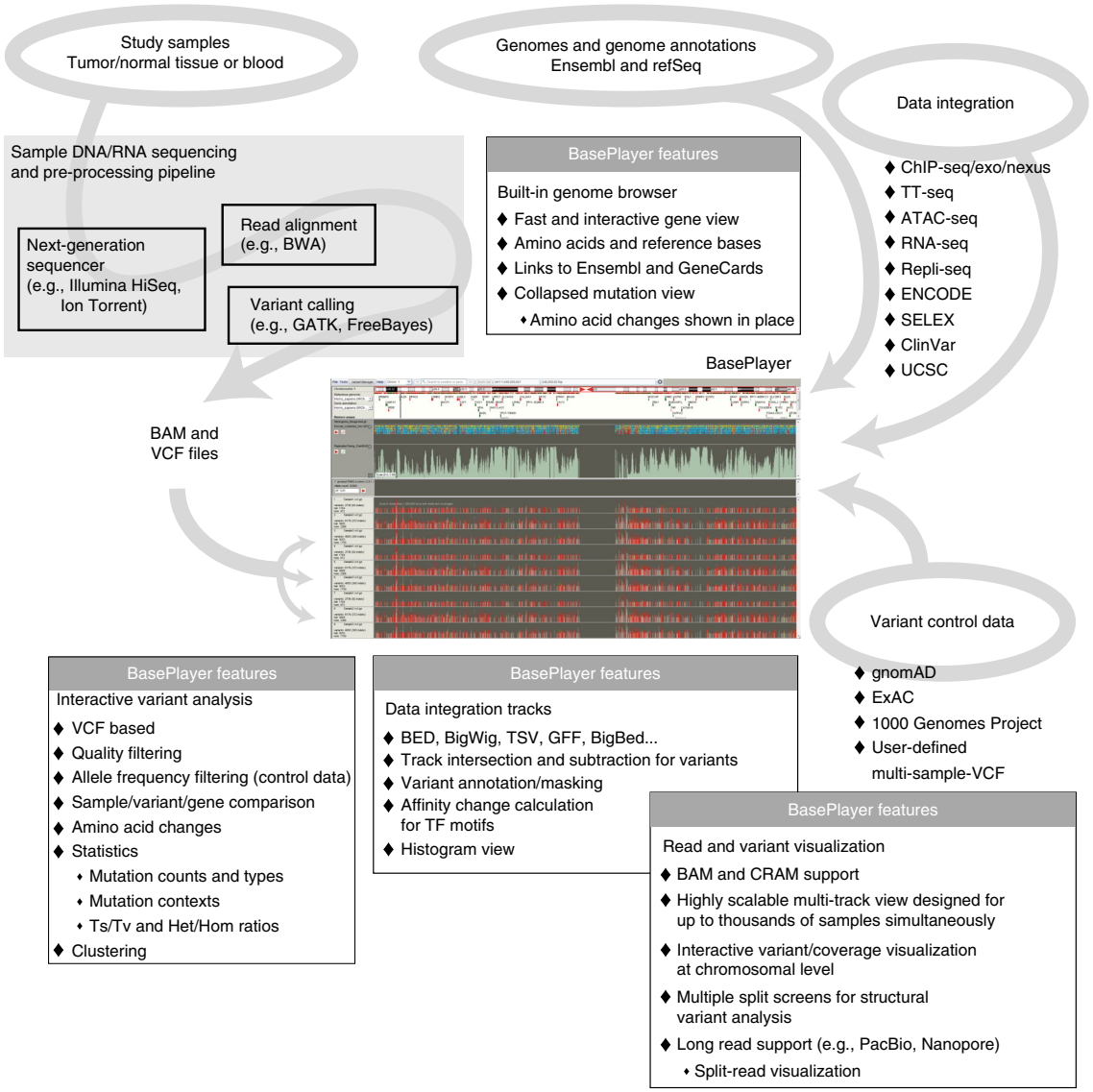

Fig. 1 | Overview of the NGS data analysis capabilities and features of BasePlayer. Sample preparation and NGS pipeline procedures, including read alignment and variant calling, are required to be performed before BasePlayer analysis. BasePlayer supports the standard file formats used in NGS data analysis, e.g., BAM and VCF files, as input file types. Ensembl reference genomes and gene annotations can be downloaded and installed automatically in the BasePlayer user interface. RefSeq reference genomes can be imported through the user interface as FASTA files and gene annotations as GFF or GTF files. Other external resources, such as ClinVar, chromatin immunoprecipitation sequencing (ChIP-seq), ATAC-seq and Repli-seq data, are available from, e.g., ENCODE and Roadmap Epigenomics can be imported into BasePlayer as annotation tracks (Step 4B(ii)). Variant control data can be obtained from, e.g., gnomAD or 1000 Genomes project websites. It is also possible to use and create a multi-sample-vcf file for this purpose. BasePlayer has a multitude of features for analysis, visualization and integration purposes listed at the bottom of the figure. ATAC-seq, assay for transposase-accessible chromatin sequencing; TT-seq, transient transcriptome sequencing.

additional processing, such as comparison of variants between samples, filtering of problematic genomic regions (e.g., repeats) and visual inspection of sequence reads, is required. In addition, the size and complexity of the NGS data in a typical study impose a substantial burden on the analysis software and user interface.

Addressing these challenges requires integrative analysis of variant and sequence data. This analysis typically consists of multiple filtering, annotation and comparison steps involving variant data (VCF files) and inspection of BAM files. To this end, we have introduced a cross-platform graphical software, BasePlayer (https://baseplayer.fi) that is designed to streamline biological discovery by enabling researchers to perform comparative genomic variant studies for human or any other species with an assembled and annotated reference genome[10–12]; see Supplementary Table 1 for a list of publications in which BasePlayer has been utilized. Distinguishing itself from other tools,

BasePlayer offers an interactive approach to variant analysis: the effect of adjusting analysis parameters is visualized by the user interface in real time, guiding parameter choices and accelerating the analysis process. Because NGS data from even a single individual can be very large—a WGS human chromosome may contain >300,000 germ-line variant calls—interactive visualization and analysis of such data require particularly efficient data structures and algorithms. With this in mind, BasePlayer is able to visualize variants of hundreds of exome-sequenced and dozens of WGS human individuals at the chromosomal level in real time due to careful design choices.

BasePlayer is well suited for the analysis of germ-line and somatic variants in both the coding and noncoding genome. Identification of rare causative germ-line variants with large effects typically takes only 10–20 min, including the software download and installation. If pedigree information is available, both coding and noncoding variants can be filtered according to the assumed pattern of inheritance, for instance, an autosomal recessive or X-linked dominant pattern, expediting causative variant discovery (Step 4A, Supplementary Tutorial 1 and Supplementary Fig. 1). To detect coding and noncoding somatic variants of relevance, for example, in cancer or clonal hematopoiesis, it is similarly quick to find genes or regions frequently mutated in the cohort, as well as clusters of variants (e.g., mutation hotspots; see Step 4B)[13]. Additional layers of data can be imported to guide the analysis further; for instance, DNA–protein interaction data can be used to screen for variants occurring at regulatory regions[11], haplotypes shared between affected individuals can be used to pinpoint an inherited mutation or third-generation long-read sequencing data (generated by, e.g., Oxford Nanopore or Pacific Biosystems instruments) can be used to allow visual inspection of the consequences of structural variation (Supplementary Fig. 2)[14]. No expertise in bioinformatics or programming is required to use the software in common use cases. A brief video introducing the software can be viewed at the BasePlayer Youtube channel ('BasePlayer Trailer', https://www.youtube.com/channel/UCywq-T7W0YPzACyB4LT7Q3g/videos).

Overview of the procedure

We demonstrate the capabilities of BasePlayer with two distinct disease genetics analysis scenarios in the Procedure (Cases 1 and 2, as described further below), both performed with an ordinary desktop computer and with 1 GB of allocated memory. Although these cases show how to replicate two previously published findings[11,15] using BasePlayer, the Procedure can be readily adapted to a wide variety of settings. In the first case, we show a procedure for finding the predisposing mutation in a family with an inherited disease (meningioma), assuming autosomal dominant inheritance, replicating a previous study described in Aavikko et al.[15] BasePlayer also supports other standard inheritance patterns, such as autosomal recessive and X-linked inheritance (Supplementary Tutorial 1 and Supplementary Fig. 1). Here, the procedure is used to find candidates for a causative, nonsynonymous mutation, which is assumed to segregate in the family, reside in a haplotype shared between cases (e.g., linkage-compatible region) and have a very low allele frequency in the general population. The dataset used in the original study included three exome-sequenced germ-line samples from the affected siblings, as well as genome-wide linkage analysis data for the same family[15]. This procedure demonstrates how to exploit BasePlayer's ability to perform between-sample variant comparison to identify variants shared between cases. The procedure also shows how gene and ROI (e.g., shared haplotypes) annotation and large-scale population variant databases can be used to narrow down the list of candidates (in the Procedure, we use the publicly available data from the gnomAD/ExAC project; see Materials)[16]. Required files for this procedure are variant (VCF) files for cases and controls, BAM files for read-level visual inspection of the data and a BED file for the genomic regions shared by affected family members (linked regions). A real-time demonstration video for this case can be viewed at the BasePlayer Youtube channel ('Familial variant study with BasePlayer', https://www.youtube.com/channel/UCywq-T7W0YPzACyB4LT7Q3g/videos).

In the second case, we analyze somatic variants in WGS cancer genomes. We describe a procedure to find somatic mutations with a possible role in tumor development by detecting recurrent mutations within genomic regulatory regions. Regulatory regions can be specified using data from, for example, the ENCODE Project[17]. This procedure replicates a study that was originally performed with somatic variant calls of 190 WGS colorectal cancers[11]. Required files for this procedure are somatic variant (VCF) files for cases, BAM files (tumor and normal WGS data) for read-level inspection, and a BED file for ENCODE regulatory and transcription factor (TF)–binding data[18] (see Materials). A real-time demonstration video of this analysis can be viewed at the BasePlayer Youtube channel ('Somatic cluster analysis demonstration with BasePlayer',

https://www.youtube.com/channel/UCywq-T7W0YPzACyB4LT7Q3g/videos). In addition to these two studies, earlier versions of BasePlayer have been used in at least 23 publications (Supplementary Table 1).

Applications of the method

Owing to an extensive supply of population-control genotypes, as well as gene and regulatory annotations available in public databases, BasePlayer has found most use in human disease genetics research settings (Supplementary Table 1). However, BasePlayer is applicable to any variant analysis setting in which an assembled and annotated reference genome is available. To facilitate nonhuman applications, Ensembl reference genomes and gene annotation files can be downloaded and used in BasePlayer directly from the user interface (Step 3).

In addition to the two procedures presented here, BasePlayer can be used in a wide variety of tasks encountered in genetics, such as population-level polymorphism analyses, mutation burden analysis for genes, mutation screening and analysis of structural variant breakpoints and target sites. For instance, genotypes obtained using non-NGS techniques (e.g., SNP arrays) can be imported and analyzed as VCF files.

The advent of third-generation sequencing (e.g., Oxford Nanopore, Pacific Biosciences) has enabled investigation of genomic regions and events for which analysis with short-read data is not feasible. Reads that are thousands of base pairs long are nontrivial to visualize, as a single read can span multiple genomic breakpoints, or multiple exons in the case of RNA data. BasePlayer natively supports long-read data. For instance, the researcher can divide the main screen between distinct genomic loci by double-clicking a read that has been mapped to multiple locations (Supplementary Fig. 2). A demonstration video on how to visualize long-read data in BasePlayer can be viewed at the BasePlayer Youtube channel ('Third-generation sequencing data visualization with BasePlayer', https://www.youtube.com/channel/UCywq-T7W0YPzACyB4LT7Q3g/videos). All split views are equally functional and contain gene annotation. This functionality can be used to, for example, search for fusion genes or targets of structural variants. In addition, BasePlayer visualizes the whole DNA fragment with respect to its split-mapped components in an inset information box (Supplementary Fig. 2).

Mitochondrial genetic diseases often show heteroplasmy, in which the disease manifests only when the abundance of mutated mtDNA molecules exceeds a certain threshold[19]. Given a cohort with mitochondrial DNA sequences, BasePlayer can be used to remove common mitochondrial variants and screen for the highly mutated allelic fraction in the remaining variants. If familial cases are available, BasePlayer can further filter for variants following a mitochondrial, or maternal, inheritance pattern to narrow down the list of candidate variants.

Comparison with other methods

BasePlayer combines features of different variant data manipulation and annotation tools, as well as visualization software packages and genome viewers into a single genetic analysis software. It incorporates annotation and filtering methods implemented in command-line tools such as Annovar, bedtools and VCFtools that were developed to manipulate and integrate VCF and BED files[20–22]. These methods include gene annotation, intersection and subtraction of genomic regions, allele frequency filtering and masking, and variant comparison. BasePlayer offers these features in a graphical interface, which reflects the parameter settings (e.g., annotations and filters in use) in real time. This allows the annotation and filtering parameters to be adjusted during the analysis without the need to create new VCF files after each adjustment of the filtering parameters, in contrast to command-line tools. Unlike many other variant analysis tools, BasePlayer has a built-in method to predict affinity changes at TF-binding sites in response to variants (see Supplementary Tutorial 2 and Supplementary Fig. 3). Core features include identification of variants shared between samples (e.g., at least five samples must share the variant) and genes that are recurrently mutated across samples (e.g., at least five samples carry a truncating variant in a gene). Furthermore, as demonstrated in Case 2 (Step 4B), BasePlayer is able to scan the genomes for clustered variants[11]. BasePlayer is also able to output the nucleotide sequence context for each variant (e.g., 5′–A[C>T]G–3′) for downstream mutation signature analysis[23].

BasePlayer tightly integrates variant analysis with data visualization, providing a low-latency built-in genome browser with separate visualization tracks for variants, sequencing data and genome annotation (provided as BEDs or BedGraphs, for example). There are several data visualization tools

and genome browsers available, such as IGV, Tablet, Artemis, Savant Genome Browser and GenomeView[24–28]. Despite our focus on genetic analysis, BasePlayer's read data and variant visualization performance is comparable to those of these browsers in terms of latency, owing to the efficient data structures and design choices made in the software (e.g., the number of split screens is limited only by the width of the monitor screen). VCF and BAM files from the same sample are combined and displayed in a single, sample-specific track; this and other design choices allow the user to view variants in hundreds of samples simultaneously. There are tools currently under development, such as seqr, VarAFT and SeqsLab (https://seqr.broadinstitute.org/, https://varaft.eu/, https://www.biorxiv.org/content/early/2017/12/27/239962, respectively), that focus on variant annotation, filtering and analysis, underlining the need for such software platforms. Unlike BasePlayer, these tools do not integrate interactive variant visualization with analysis and seem to currently support only human analyses, mainly focusing on the coding regions, limiting their applicability. Furthermore, tools have been developed for specific tasks in which variant and sequence data must be integrated. One example is Viper, which allows visual validation of variants against NGS data[29]. BasePlayer improves on Viper by allowing more than one BAM file to be examined at the same time at the same locus interactively, together with genomic annotation data, a critical feature when validating cancer somatic variants, for example (see Case 2, Step 4B(ix) and https://baseplayer.fi/BPmanual/content/ui.html#tbrowser). In many external server or cloud-based analysis platforms, such as BaseSpace (https://basespace.illumina.com/), SeqsLab and Chipster[30], the data reside in the cloud, complicating analysis of sensitive or massive data. To address such scenarios, BasePlayer is fully portable, and neither an active network connection nor registration is needed. Mutation-ranking tools such as OncodriveFML, Genomiser/Exomiser and FunSeq2 are designed to score and annotate the disease-causing potential of variants in the human genome[31–33]. These tools can be useful for discovering potential causative mutations. Although BasePlayer does not implement a variant-ranking functionality, annotations from external tools can be imported as annotation tracks and used to filter variants. Moreover, BasePlayer is able to output variants in a format accepted by OncodriveFML as input[31].

BasePlayer's flexible and interactive user interface facilitates rapid discovery of candidate causative mutations from variant data. It also allows large-scale comparative genetic analyses to be performed with an ordinary desktop computer by a single person with basic computing skills. To facilitate disease genetics, BasePlayer supports analysis under single-gene inheritance patterns such as autosomal dominant and recessive patterns in familial data (see Step 4A(v), Supplementary Tutorial 1 and Supplementary Fig. 1). Pedigree-based analysis is also available in seqr (Broad). Whereas seqr focuses on human familial diseases, BasePlayer also supports other species and is not restricted to familial analyses.

Limitations

BasePlayer does not annotate the variants with predicted functional impact on the protein (e.g., neutral or damaging). However, it is possible to create a VCF file in BasePlayer that contains the results of BasePlayer analysis, and to analyze the variants further in Ensembl Variant Effect Predictor or any variant-impact prediction tool that accepts VCF input. In addition, annotations from public databases or tools, such as dbSNP, ClinVar, ICGC, COSMIC and PubMed, and pathogenicity predictors can be viewed at the VarSome website (https://varsome.com/) by clicking the variant in the 'result table' or in the sample track (see Procedure, Case 1). Alternatively, it is possible to filter or annotate variants in BasePlayer by their predicted impact with annotation tracks prepared for this purpose. Impact scores from three recent databases can be used to score variants by their predicted impact: CADD, DANN and M-CAP[34–36]. See Supplementary Tutorial 3 and Supplementary Fig. 4 for instructions on how to install these resources.

BasePlayer allows interactive visualization of all variants in all samples on a single chromosome at once. However, this ability is the most CPU- and memory-intensive feature of the software and may cause interactive usage to appear sluggish when a large number of samples are opened at the same time, depending on the available CPU and memory resources. In our use, we routinely analyze hundreds of WGS samples in BasePlayer using 1.2 GB (maximum memory allocation in 32-bit systems) of allocated memory. To analyze a larger dataset, more memory must be allocated to the software. For instance, BasePlayer consumes ~2.3 GB of memory when ~28 million somatic single-nucleotide variants (SNVs) in 2,700 WGS cancers have been loaded. If the memory consumption is excessive, BasePlayer disables the variant visualization but is still able to analyze and annotate variants in batches (default window size = 1 Mbp) and write the annotated results directly to a file.

BasePlayer's approach to filtering candidate variants by population genotypes, inheritance patterns and layers of domain-specific data facilitates the discovery of rare causative variants, especially those with large effects. Numerous methods for rare-variant association testing have been developed to identify predisposing variants when effect sizes are expected to be smaller (e.g., SKAT-O[37]). Although BasePlayer does not yet directly implement such techniques, it can still be used as a quality control, and to filter and annotate variants before carrying out any association test that supports VCF input format. Conversely, candidate variants highlighted by an association test can be imported into BasePlayer to filter, annotate and visually inspect candidates together with additional layers of data. In summary, analyses conducted using statistical techniques or BasePlayer are not mutually exclusive but can instead complement each other.

Finally, because BasePlayer is operated through a graphical user interface and does not currently offer an application programming interface, it cannot be included as a part of automated analysis pipelines or scripts, unlike command-line tools.

Experimental design

Input data formats

A typical BasePlayer session involves matched NGS data and variant calls together with control genotypes and genomic regions, which can be used to filter the candidate variants (Fig. 1). Standard VCF files compressed with bgzip and indexed with Tabix are required for variant analysis[38]. VCF files are created by many types of variant-calling and annotation software, including GATK HaplotypeCaller, Torrent Variant Caller plug-in (IonTorrent) and Mutect[8,9]. To visualize read sequence data and sequencing coverage, indexed BAM files are required. BAM files are created by short-read aligners such as BWA or Bowtie, or by NGS workflows such as GATK[9,39,40]. We recommend preparing a pair of VCF and BAM files for each sample (case) to be analyzed, giving them an identical file name prefix and placing them in the same folder together with the respective index files (i.e., the sample folder should contain samplename.vcf.gz, samplename.vcf.gz.tbi, samplename.bam and samplename.bam.bai). With this setup, the VCF and BAM files from the same sample are matched, and read data are visualized alongside the variants in the sample track. In Unix environments, file links can be used to arrange data located in multiple folders to be accessible from a single folder.

Variant filtering

To filter variants that are common in the population, it is possible to use reference variant data in BasePlayer. Set the allele frequency threshold to reflect the characteristics of the studied disease, such as incidence rate. We provide a snapshot of the Genome Aggregation Database (gnomAD) for the human genome GRCh37 containing whole-genome variants of $n = 15{,}496$ and whole-exome variants of $n = 123{,}136$ individuals (see Downloads at https://baseplayer.fi). To reduce processing and download times, as well as file sizes, we removed all other information (e.g., gene annotations) except allele frequencies from the gnomAD VCF files and merged all remaining data into a single file. This preprocessing is for convenience only: gnomAD data available at http://gnomad.broadinstitute.org/downloads can be used in BasePlayer without any preprocessing steps. It is also possible to create user-defined control files (multi-sample VCFs) in BasePlayer as follows: open multiple control sample VCF files (e.g., samples from healthy family members or general population) and set the variant filter thresholds to zero to include low-quality variant calls in the control file. This strategy allows for effective filtering of technical artifacts in the data, otherwise abundant in regions that are difficult to sequence. Then, annotate all variants and write the results to a VCF file. The resulting VCF file can be opened as a control track (see Step 4A(iii) for more details).

In the case that only a subset of genes are of interest, create a text file (.txt) containing names or ENSG (Ensembl gene identifiers) codes for the desired genes, add the file as an annotation track and apply the track to exclude variants outside the genes of interest (Step 4A(ii)).

Quality-filtering thresholds (e.g., minimum (min.) coverage, min. quality score and min. allelic fraction) depend on the sequence data type and quality. Lower thresholds increase filtering sensitivity, but lower specificity. For instance, the allelic fraction threshold can be set higher in germ-line variant analysis (e.g., 20%) than in analysis of subclonal somatic variation (e.g., 5%).

Data integration

Genomic data other than sequencing or variant data can be opened in BasePlayer as tracks. BasePlayer supports standard file formats, including BED, BedGraph, bigWig, bigBed, GFF and

tab-separated (TSV) formats. This allows integration of data from various experiments and analyses into variant analysis, for instance, peak calls from DNAse-seq, MNAse-seq and ChIP-seq experiments; methylation levels from whole-genome bisulfite sequencing or methylation array experiments; regions spanned by copy-number and structural variants; replication timing (e.g., Repli-seq); chromatin segmentation (e.g., ChromHMM); repeats (e.g., RepeatMasker); mappability; and genomic loci of specific interest (e.g., known pathogenic variants in ClinVar, candidate genes). These data can then be used to annotate and filter variants and, together with NGS data and variants, to provide a unified view of the genomic landscape of the study subjects.

Variant annotations

Amino acid changes are calculated, and gene information is added to variants using indexed reference sequence (FASTA) and gene annotation (GFF3), which are loaded by default when BasePlayer is launched. The gene annotation file contains, among other things, chromosomal positions and codon phases for all protein coding exons, which are used to fetch codon sequence triplets from the reference sequence file. An amino acid change is derived from the variant position and the base change (reported in the VCF file) in the fetched codon sequence. BasePlayer annotates synonymous, nonsynonymous, nonsense, splice-site, frameshift, inframe, untranslated region (UTR), and intronic and intergenic variants, and is compatible with Annovar annotation[20]. Adjacent genes are reported for intergenic variants.

Materials

Equipment

- A modern computer running Windows, Linux or macOS operating system with Java Runtime Environment installed (64-bit recommended; http://www.oracle.com/technetwork/java/javase/downloads/jre10-downloads-4417026.html)
- Internet connection (required for the installation of new reference genomes)
- Mouse with at least two buttons (recommended for smooth genome navigation)
- BasePlayer, available free of charge at https://baseplayer.fi

Input data files for both example cases

- The human reference genome (GRCh37) and gene annotation data for both example cases can be downloaded from the Ensembl database (Release 87) as in Step 3.

Input data files for Case 1

- VCF and BAM files from any exome- or genome-sequenced patients with a suspected inherited disease. The example analysis was carried out with exome-sequencing data from meningioma patients (Case 1), described in Aavikko et al.[15]. The publicly available data used in Supplementary Tutorial 1 can also be used to test this case. Download the data from https://baseplayer.fi/BPmanual/content/analysis.html#inheritance. ▲ **CRITICAL** Note that these VCF files have the quality score (QUAL field) fixed to 50 for all variants; thus it is not possible to use quality score as a filtering parameter when using this particular sample set.
- The exome control files (both gnomAD_exomes_ALL_GRCh37.vcf.gz and the .tbi index) can be downloaded from https://baseplayer.fi/controls/. Initially, GnomAD variant frequency data were downloaded via the gnomAD website (http://gnomad.broadinstitute.org/downloads)[16].
- M-CAP pathogenicity prediction data, described in Jagadeesh et al.[36]. The file 'mcap_v1_0_forBP.bed.gz' can be downloaded from https://baseplayer.fi/tracks/.

Input data files for Case 2

- WGS data for colorectal cancer samples and identification of somatic variants (Case 2) described in Katainen et al.[11]. European Genome-Phenome Archive (EGA) accession no. EGAS00001003010.
- ENCODE data for regulatory regions as published in Hoffman et al.[41] ▲ **CRITICAL** We merged the data from six different cell lines (GM12878, H1-hESC, HeLa S3, Hep G2, HUVEC and K562) to create the ENCODE annotation file.
- Position-specific matrices for TFs obtained from Jolma et al.[18] and the JASPAR database[42]. TF-binding sites for GRCh37, aligned with MOODS[43], and the ENCODE annotation file. Download from the BasePlayer website (BasePlayer Download page at https://baseplayer.fi/downloads.html, under 'Annotation tracks').

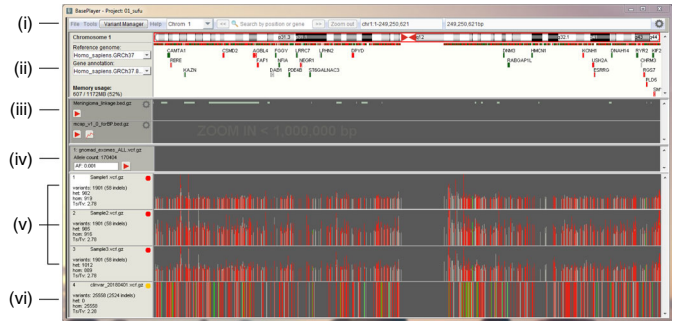

Fig. 2 | The main window of BasePlayer, displaying three samples, a genomic region track and a population control data track. Variant view of chromosome 10. (i) Toolbar contains tools for managing the data and navigating the genome. (ii) Genome bar visualizes chromosome bands and genes. Reference sequence and gene annotation can be changed from dropdown menus on the left. Memory usage shows the used and allocated memory for BasePlayer. (iii) Region tracks are used to exclude and annotate variants. In addition, various types of regional- or base-specific scores can be visualized as histograms and TF-binding motifs can be visualized as sequence logos. In this case, tracks containing linkage-compatible regions and M-CAP annotations are opened. Clicking the red 'play' button will apply the track; this will, by default, exclude any variants outside the region and annotate visible variants. See Case 2 (Step 4B(ii) and (iii)) and Supplementary Tutorials 2, 3 and 4 for more track options. (iv) Control tracks can be used to exclude common polymorphisms by setting the allele frequency threshold such that variants whose allele frequencies in the control data exceed the threshold are removed. (v) VCF and BAM files of the sample can be overlaid for visualization. Sample name and statistics for currently visible variants are shown on the left. Vertical lines represent variant calls in a sample, colored red, green (SNVs or indels in coding region, respectively) and gray (variants in noncoding region). Height is relative to the sequencing coverage at the variant locus. (vi) Sample VCF track is set as an annotation track. ClinVar data are opened in this case.

Additional Java packages

- Htsjdk (v.1.141, https://github.com/samtools/htsjdk). BasePlayer uses Htsjdk index readers for BAM, CRAM and Tabix indexed files. CRAM reader was optimized for BasePlayer by modifying CRAMFileReader and related classes from the Htsjdk package.
- Index readers for BigBed and BigWig files. Download from https://github.com/lindenb/bigwig; these were originally written by M. Decautis and J. Robinson for the Integrative Genomics Viewer (Broad Institute)[24].

Procedure

Download and install BasePlayer ● Timing 5 min

1 Download BasePlayer from https://baseplayer.fi/downloads.html. Click the operating system-specific installation package to start the download. Launch the installer and follow the instructions. In case of insufficient rights for your desktop computer to launch executable files, download the portable multi-platform JAR package ('BasePlayer portable package') from the downloads page and decompress it to a folder with writing permissions. To test the features of BasePlayer discussed in this Procedure with example NGS data, download the 'BasePlayer portable package with example data' from the downloads page or optionally family trio data (download links at https://baseplayer.fi/BPmanual/content/analysis.html#inheritance).

▲ **CRITICAL STEP** Java JRE v.1.8 or a newer version is required to run BasePlayer. This procedure was conducted with the portable JAR package and 64-bit Java Runtime Environment in a Windows 7 system.

? TROUBLESHOOTING

2 To start BasePlayer, click the BasePlayer executable (or optionally Launcher.jar with execute permissions, if the portable package is used). If an updated version of the software is available (green update button in 'File' menu), click 'File' > 'Update' to download the updated version and restart the software (Fig. 2,i).

! CAUTION If the dataset to be analyzed contains >10 million variant calls, we recommend allocating ≥2 GB of memory to the Java Virtual Machine for optimal usage. To increase memory allocation, edit the BasePlayer.vmoptions file in the BasePlayer install folder. Instructions can be found in the vmoptions file. When using the portable package and launching BasePlayer with the Launcher.jar file, edit the file 'config_baseplayer.txt' in the user home directory of the operating

system, changing the memory limit to, e.g., 4 GB (make sure that the 64-bit Java Runtime Environment is installed on the system), and restart the program with Launcher.jar.

? TROUBLESHOOTING

3 When BasePlayer is launched for the first time, a genome selector menu appears (note: if 'BasePlayer portable package with example data' was selected in Step 1, a reference genome is already installed, and the genome selector menu does not appear). Select the preferred reference genome and click 'Download'. Revisit this menu by clicking the 'Add reference genome…' dropdown item on the left in the genome bar (Fig. 2,ii). If storing the downloaded genome is not possible in the default location due to restricted permissions, select a folder in which to store the genome file, when prompted.

! CAUTION Note that the example data provided in the 'BasePlayer portable package with example data' contain only the reference sequence and gene annotation for human chromosome 20. Furthermore, example samples are from healthy, unrelated individuals, and thus the results of Cases 1 and 2 in this procedure cannot be replicated with the example data as such. Instead, the example data can be used to perform all the steps in this Procedure, even though the end result will be different from what is shown in these procedures.

? TROUBLESHOOTING

NGS data analysis using BasePlayer

4 Use BasePlayer to analyze the NGS data by following option A to use Case 1 example data or option B to use Case 2 example data.

(A) **Case 1: finding a dominant pathogenic inherited variant using exome-sequencing data**

● **Timing 5 min**

(i) Load the germ-line variant data by clicking 'File' > 'Add VCFs' and selecting the VCF.gz files corresponding to cases (press 'Ctrl'/'Cmd' or 'Shift' to select multiple files simultaneously). Note that BAM files are opened automatically for the same samples, if they are appropriately named and found in the same folder as the VCF files (Experimental design). A chromosome-level variant view is now shown, with each sample displayed on a separate track (Fig. 2).

? TROUBLESHOOTING

(ii) Annotate the variants using external resources. Open additional track files by selecting 'File' > 'Add tracks' to use in variant filtering and/or annotation. Many of the standard file formats are accepted (e.g., BED and bigWig). New tracks will be shown under the genome bar (Fig. 2,iii). In our example, the track file (in BED format) contains linkage-compatible regions for the studied family. Click the 'play' button on the left sidebar of the track to remove variants outside the specified regions (intersect; default function). The intermediate results after each filtering step are shown in more detail in Supplementary Fig. 5. To predict the possible pathogenicity of variants, add the mcap_v1_0_forBP.bed.gz file as a new track. The file (and the index) can be downloaded from https://baseplayer.fi/tracks/. Click the 'play' button on the left sidebar of the track to annotate the variants with the M-CAP scores (see Supplementary Tutorial 4 and Supplementary Fig. 6 for the recommended usage of the file). We have set the M-CAP score limit to 0.025 (recommended by the authors of the method) in this case. To download and add the ClinVar annotation, see Supplementary Tutorial 5 and Supplementary Fig. 7.

▲ **CRITICAL STEP** The resources mentioned here work as an example, but users are not limited to those. For instance, Table Browser in the UCSC Genome Browser (https://genome.ucsc.edu/) is a comprehensive source for useful tracks that can be imported into BasePlayer (see the instruction video 'Loading external annotation resources to BasePlayer' in our Youtube channel: https://youtu.be/nSQX2Ctwhso).

? TROUBLESHOOTING

(iii) Open a VCF file containing population variants by selecting 'File' > 'Add controls', which opens a new track on top of the sample tracks (Fig. 2,iv). Here, we use an exome-specific control genotypes file extracted from gnomAD[16]. The exome control file (gnomAD_exomes_ALL_GRCh37.vcf.gz and the .tbi index file) can be downloaded from https://baseplayer.fi/controls/ or from http://gnomad.broadinstitute.org/downloads ('All chromosomes sites VCF' under 'Exome Data'). First, set the allele frequency threshold (see 'Experimental design' for further information) in the text field on the left sidebar (0.001 in this case).

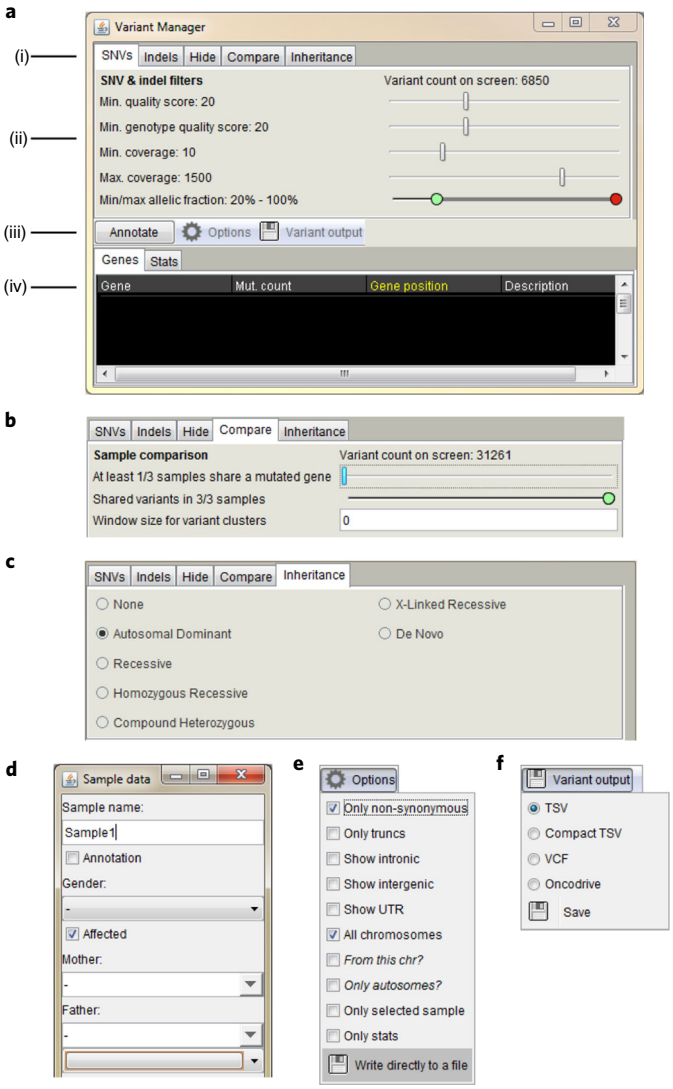

Fig. 3 | Variant Manager user interface and functions. a, (i) Panels for variant quality filtering, variant hiding and sample-wise comparison. (ii) SNV and indel quality filters. Separate filters for indels can be activated in the 'Indels' tab; the same filtering is used by default for SNVs and indels. (iii) Variant annotation panel, including gene effect and output writing options. (iv) Result table shows a list of genes and variants that pass the filters set in the Variant Manager, as well as any activated track. This view also includes tables for variant statistics and region track annotations. **b**, Sample-wise comparison panel. The upper slider sets the minimum number of cases that must harbor a variant in the same gene for these variants to be shown. It can thus be used if the sample set consists of cases with a common disease, which might not share the same specific variant even though the same gene is mutated (e.g., sporadic cases or somatic mutations in cancer). The lower slider compares individual variants. For instance, the value 2/3 means that only variants shared by two or three (all) samples are included and visible on the screen. 'Window size for variant clusters' sets the variant clustering behavior and will be demonstrated in Case 2 (Procedure). **c**, Inheritance pattern selector. **d**, Sample data dialog, in which the user can change/set sample name, gender, disease state and parents. **e**, Variant annotation options, which can be used to select the desired functional classes and options for variants, which are included in the results. **f**, Output selector for variant annotation results. Here, the user can select the desired output format for the result file.

Click the 'play' button to apply the control track. Variants appearing in the control file with a higher frequency than the specified value are then excluded from the analysis and removed from the screen.

(iv) To filter, compare and annotate the variants, click 'Variant Manager' on the toolbar (Fig. 3). Use the sliders to set quality and coverage thresholds for the variants (Fig. 3a,ii). These values depend on the data quality and research setting; we typically use values 20, 10 and 20% for the variant quality, coverage and allelic fraction thresholds, respectively

(see 'Experimental design' for further information), when analyzing germ-line data. Note that 'Min. quality score' refers to the QUAL (Phred scale) field and 'Min. genotype quality score' refers to the genotype quality (GQ) field in the VCF file. Quality values are given in Phred scores and shown as reported by the variant caller that generated the VCF files (GATK HaplotypeCaller in this example). Homozygous variants are filtered out either by setting the maximum allelic fraction to, e.g., 95%, by using the red handle of the 'Min/max allelic fraction' slider (Fig. 3a,ii) or by selecting the 'Hide homozygotes' checkbox in the 'Hide' tab of 'Variant Manager'.

(v) Variant comparison between samples can be performed either by comparing variants or mutated genes without prior knowledge of sample relations or by using inheritance patterns of related samples and known disease statuses. The former option is commonly used in sporadic and somatic settings or in any custom variant-management tasks, whereas the latter approach is used in familial cases, when affected and/or unaffected relatives are included in the sample set. In this example, the study can be conducted both ways, as all individuals are affected and there are no parents available for the siblings. The 'Compare' tab in Variant Manager contains sliders for gene- and variant-level comparisons, which are used to detect shared/unique variants, variant clusters and commonly mutated genes in studied samples, regardless of the relations or disease statuses (Fig. 3b). The 'Inheritance' tab contains options for inheritance patterns, which are applied in annotation to output only variants that are compatible with the selected pattern (Fig. 3c). To find variants that are shared among *n* cases without assumed inheritance patterns, click the 'Compare' tab in Variant Manager and set the value of the first slider to *n* (Fig. 3b). Because all samples used in this example are first-degree relatives and share the disease, the 'Shared variants' slider is set to 3/3 (i.e., only variants shared by all samples are included and visible on the screen).

To apply inheritance patterns in variant annotation, at least one individual must be set as affected. This is done in the 'Sample data' dialog (Fig. 3d), which appears by left-clicking the sample name or right-clicking the left sidebar on sample tracks. In this case, all samples are set as affected. The inheritance pattern, based on the segregation of phenotype in the family, can be selected in the 'Inheritance' tab of Variant Manager (Fig. 3c). In this case, the segregation suggests an autosomal dominant inheritance pattern, which is selected.

(vi) Annotate the candidate variants with gene information. At this point, variants are included if they are heterozygous, of sufficient quality, overlap linkage-compatible regions, are very rare in the population, have an M-CAP score >0.025 and are shared by all cases or are compatible with an autosomal dominant inheritance pattern (as determined in Step 4A(v)). In this example, we are interested in nonsynonymous variants only and variants that appear in any of the chromosomes: first click the 'Options' button (Fig. 3a, iii) and then the 'Only non-synonymous' and 'All chromosomes' checkboxes (Fig. 3e). Begin the annotation by clicking the 'Annotate' button. This will result in BasePlayer analyzing each of the chromosomes and displaying the annotated variants in the 'Genes' tab (Fig. 3a,iii,iv). Using the settings, controls and annotations selected above, the resulting candidate list consisted of five mutations in five genes, of which the variant in *SUFU* had highest M-CAP score and was not present in the gnomAD controls; it was subsequently found to be the prime candidate meningioma-predisposing mutation in the family[15] (Fig. 4a).

(vii) Write the list of candidate variants to a file by clicking the 'Variant output' tab, selecting the preferred file format and clicking 'Save'. The output list can be written in tab-separated (TSV), VCF or OncodriveFML-compatible format (Fig. 3f) (more information is given in the 'Anticipated results' section).

(viii) *Read-level inspection of variants*. Variants can be visualized and inspected in the sequencing read data (if BAM files were opened in Step 1) interactively by double-clicking the variant or gene in the result table (Fig. 4a). When the sample row is double-clicked in the result table, BasePlayer expands the selected sample track and zooms in to the variant locus (Fig. 4b). Reads will appear if a BAM file is available for the VCF file. If you double-click the 'Multiple' row in the result table, multiple samples are shown simultaneously. You can expand and shrink the tracks manually by dragging the mouse to the top-right and bottom-left corners (see 'Browsing genome and samples with BasePlayer' in our Youtube channel: https://youtu.be/EJHZXLf6nQM).

(ix) *Inspecting external resource annotation of variants*. Data from annotation tracks can be inspected in the result table either by selecting the corresponding tab or by clicking the

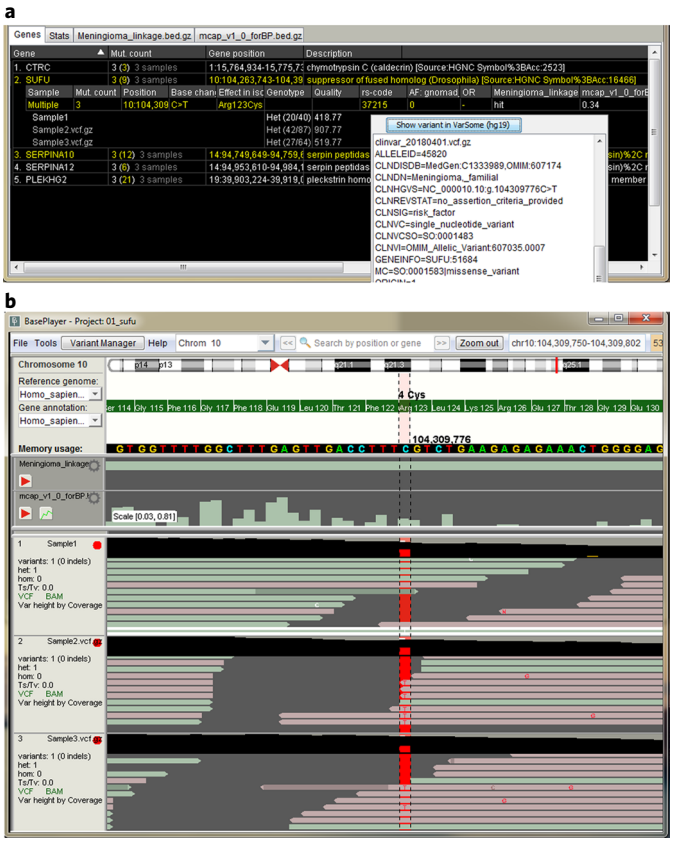

Fig. 4 | Candidate genes in the result table and variant visualization. a, Candidate gene list in the result table of the Variant Manager. The 'Mut. count' column shows the total mutation count and the number of samples having a mutation in the gene. Colored numbers in parentheses indicate mutation counts and effects in all isoforms; e.g., if a gene has four isoforms overlapping three missense mutations, then the number 12 will be shown (in yellow) inside the parentheses. The colors red, yellow, green, light gray and dark gray correspond to the mutation types frameshift/nonsense, missense/inframe indel, synonymous, UTR and intronic/intergenic, respectively. Mutation information can be expanded by clicking the gene name, which displays the genomic position, amino acid change, allele frequency in the control file and enabled annotation tracks. Clicking the variant shows a VCF information pop-up, which includes a link to the VarSome website and annotation VCF track data (ClinVar in this case). Double-clicking the gene or mutation will zoom in the sample track view to the selected locus. **b**, Sequence-level zoom of the *SUFU* mutation. The VCF information pop-up can also be opened by clicking the variant in the sample track (vertical red lines).

gene (Fig. 4a). For instance, the selected variant in Fig. 4a has an M-CAP score of 0.34, overlaps the linkage region and has an allele frequency of 0 in the gnomAD control file. Click the variant for more information (VCF information pop-up). ClinVar data are shown for this variant alongside the VCF information. Click the 'Show variant in VarSome (hg19)' to open the VarSome website (https://varsome.com) for the variant locus (registration may be needed). VarSome lists and visualizes a number of additional annotations from multiple sources, such as dbSNP, ClinVar, ICGC, Cosmic and PubMed, as well as multiple pathogenicity predictors.

(B) **Case 2: somatic mutation clusters in the regulatory genome** ● Timing **10 min**

(i) Load the somatic variant data by clicking 'File' > 'Add Samples' and selecting the VCF.gz files. BAM files are opened, if available, automatically for the same samples as used in Case 1 in Step 4A(i).

? TROUBLESHOOTING

(ii) Add a track file for regulatory regions by clicking 'File' > 'Add tracks'. In this example, a file containing regulatory annotations defined by the ENCODE Project is used. This file can be downloaded from https://baseplayer.fi/tracks/. Variants and the ENCODE regulatory regions are now loaded and appear on the screen (Fig. 5).

? TROUBLESHOOTING

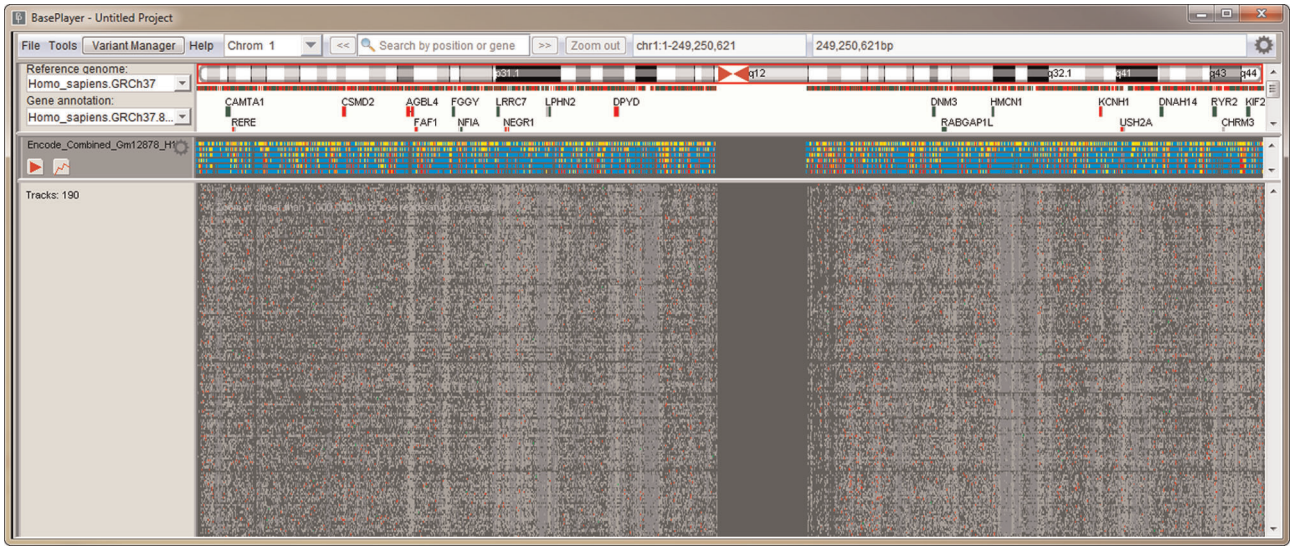

Fig. 5 | Somatic variants in the regulatory genome. All somatic single-nucleotide variants of 190 colorectal cancers (WGS) in chromosome 1 are visualized. Additional track in the middle visualizes ENCODE regulatory regions, color-coded as: blue, CCCTC-binding factor (CTCF) binding sites; dark yellow, enhancers; light yellow, weak enhancers; red, transcription start sites; and light red, promoter flanks.

(iii) Add another track file, which contains predicted or measured binding sites of TFs relevant to the disease under investigation. In this example, we use binding sites for hundreds of TFs contained in the file TFbinding_sites_SELEX_GRCh37.bed.gz (track and index files downloadable from https://baseplayer.fi/tracks/)[18].

(iv) Open Variant Manager and adjust the preferred filtering thresholds (see 'Experimental design' for further information) for your samples (Fig. 6a). Depending on the estimated tumor purity of the samples, minimum allelic fraction should be set accordingly; the settings used here are shown in Fig. 6. In this example, we exclude indels from the analysis by clicking 'Hide indels' under the 'Hide' tab, because indels at low-complexity regions (e.g., microsatellites) would otherwise cause an excessive amount of nondesired variant clusters genome-wide (Fig. 6b).

(v) Adjust the variant clustering settings in the 'Compare' panel (Fig. 6c). Set the window size for variant clusters at the bottom of the compare panel (50 bp in this example); this is the maximum distance between consecutive clustered variants. A cluster is defined as a set of variants in which the distance between any two adjacent variants does not exceed the window size. Note that the cluster width (distance between the left-most and right-most variant) can thus exceed the window size. The 'Common variants' slider is used to set the minimum number of variants in a cluster (Fig. 6c).

(vi) Adjust the settings for the regulatory region track by clicking the cogwheel symbol in the left sidebar of the track (Fig. 6d). In this example, we use the default options. When the 'Intersect' option (the default) is selected, variants outside the regions specified by the track are excluded and hidden on the screen. When the 'Subtract' option is selected, variants inside the region are excluded. After the options have been set, apply the track by clicking the track's play button.

(vii) At this point, somatic SNVs that overlap the ENCODE regulatory regions and belong to a variant cluster consisting of at least four variants are included and visible. In each variant cluster, any two consecutive variants are at most 50 bp apart. Click 'Show intronic', 'Show intergenic' and 'Show UTR' to include all noncoding variants in the results (Fig. 6e). Scan all mutation clusters genome-wide by clicking 'Annotate' in Variant Manager.

(viii) The results are listed in the result table of the 'Clusters' tab in Variant Manager (Fig. 7a). Double-click a cluster or individual variant to jump to the genomic location of the cluster or variant (Fig. 7b). If TF-binding data are added as a track (Step 4B(ii)), then TF-binding motifs are shown in the track when zoomed in to nucleotide resolution (see Supplementary Tutorial 2 and Supplementary Fig. 3 for affinity change annotation using TF-binding data).

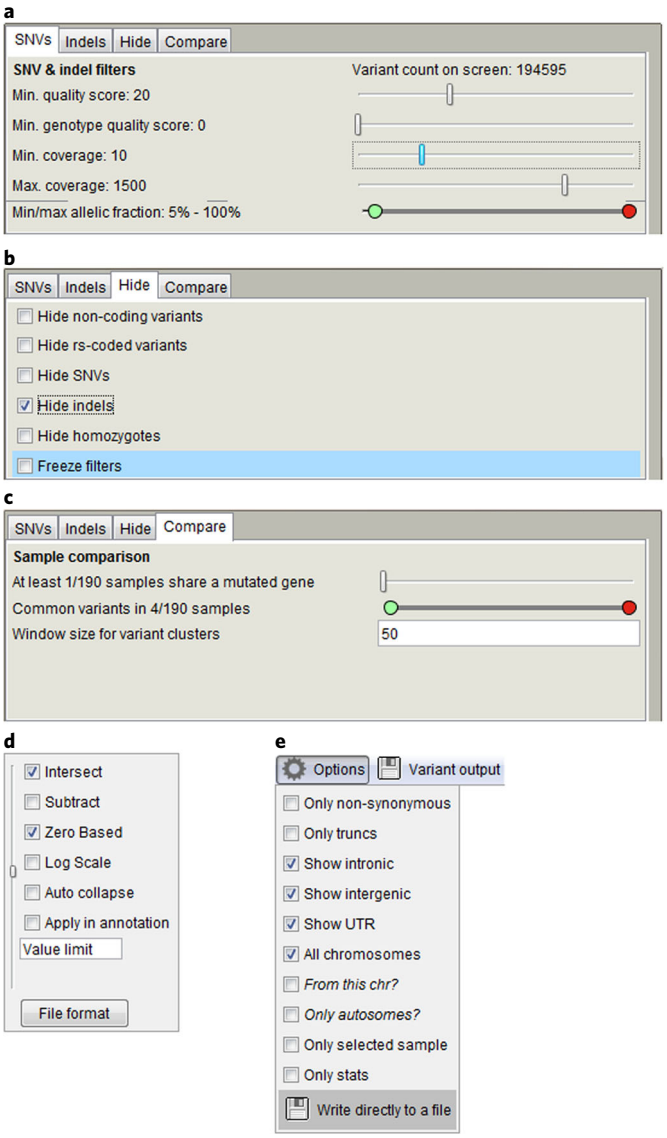

Fig. 6 | Variant Manager settings in somatic cluster analysis. a, Example of filtering settings for somatic variants. 'Min. allelic fraction' is set to as low as 5% to include subclonal variants in the cluster analysis. **b**, Indels are excluded from the analysis, as they produce an excessive amount of clusters at low-complexity regions of the genome. **c**, Window size is set to 50 bp, which determines the maximum distance between two adjacent variants in a cluster. The shared variant slider is used to exclude any cluster with fewer than four variants. **d**, Default values of annotation track options. 'Intersect' and 'Subtract' options are mutually exclusive. If both are unchecked, variants are still annotated when the track is applied. **e**, With these settings, all noncoding mutations in all chromosomes will be annotated.

(ix) *Visual validation of variants*. Variant-quality metrics in VCF files are provided by the used variant caller. Although the false-positive rates of variant callers have decreased as methods have improved, read-level inspection of variant calls and flanking regions is still often required. In particular, low-allelic-fraction somatic mutation calls in the noncoding genome, as in this case, should be inspected visually. The integrated variant analysis and visualization in BasePlayer enables rapid workflow for variant validation; double-click the specific sample in the result table to zoom in to the variant locus and expand the sample track. Reads appear on the sample track (Fig. 8) if a BAM/CRAM file is opened alongside the VCF file (Step 4A(i) and (vii)). Using read-level zoom, inspect the qualities, orientation and mismatch rates of the overlapping and surrounding reads. In more ambiguous cases, read-level inspection among multiple samples can be

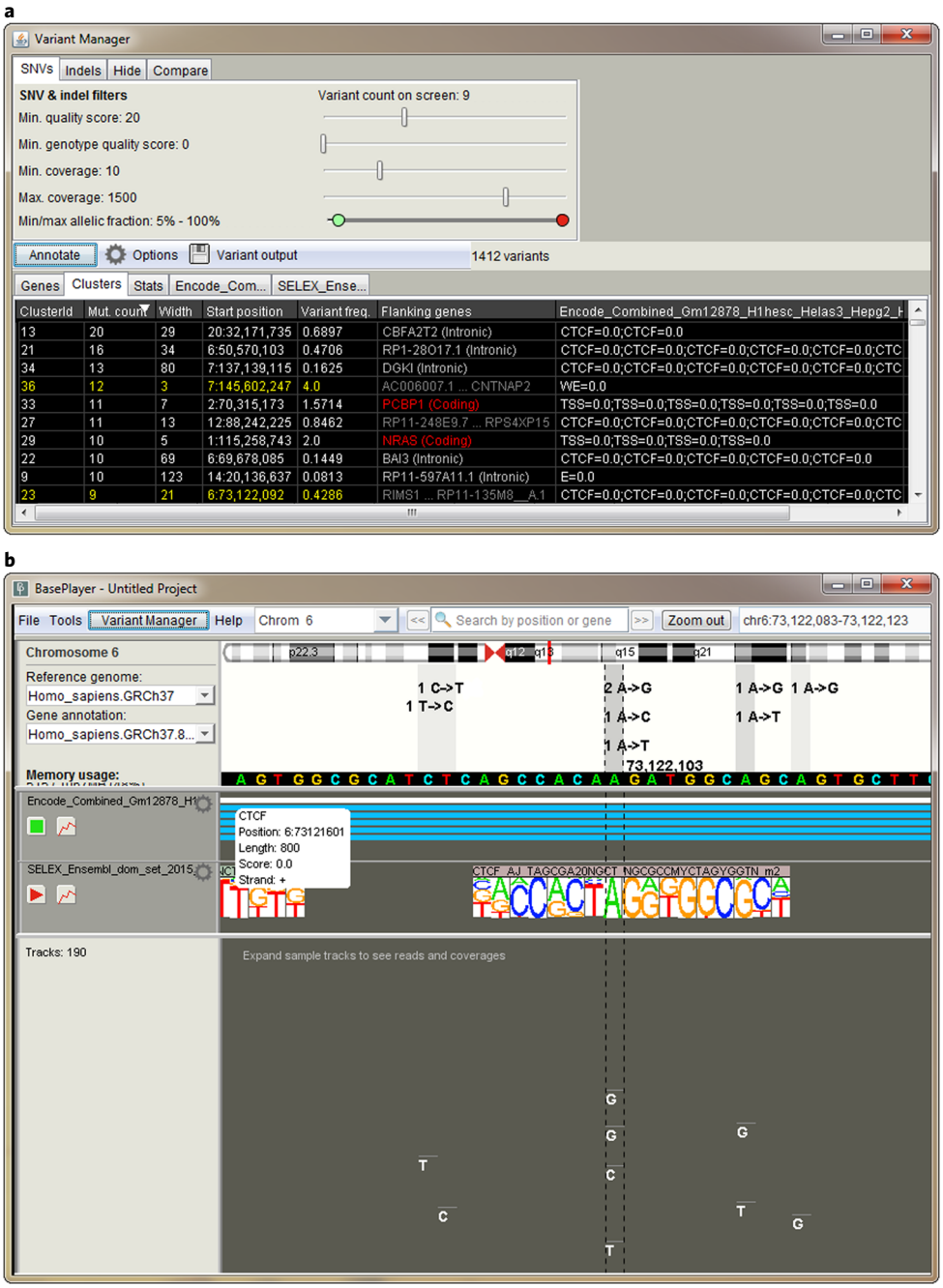

Fig. 7 | Somatic clustering results. a, The 'Clusters' tab shows a list of all annotated mutation clusters. **b**, Motif visualization with JASPAR and SELEX data. Closer inspection of the cluster highlighted in Katainen et al.[11] shows mutation hotspots at an occurrence of the CTCF-binding motif.

useful; you can visualize reads from multiple samples easily in BasePlayer by scrolling through the samples and shrinking the sample tracks (see example in Fig. 4b). In addition, the reference sequence and annotation tracks, such as the 'repeat masker' and 'mappability' tracks, provide additional information about variant calls when inspecting variant quality.

(x) Write the cluster analysis results to a TSV file by clicking 'Variant output' and 'Save' in Variant Manager.

? TROUBLESHOOTING

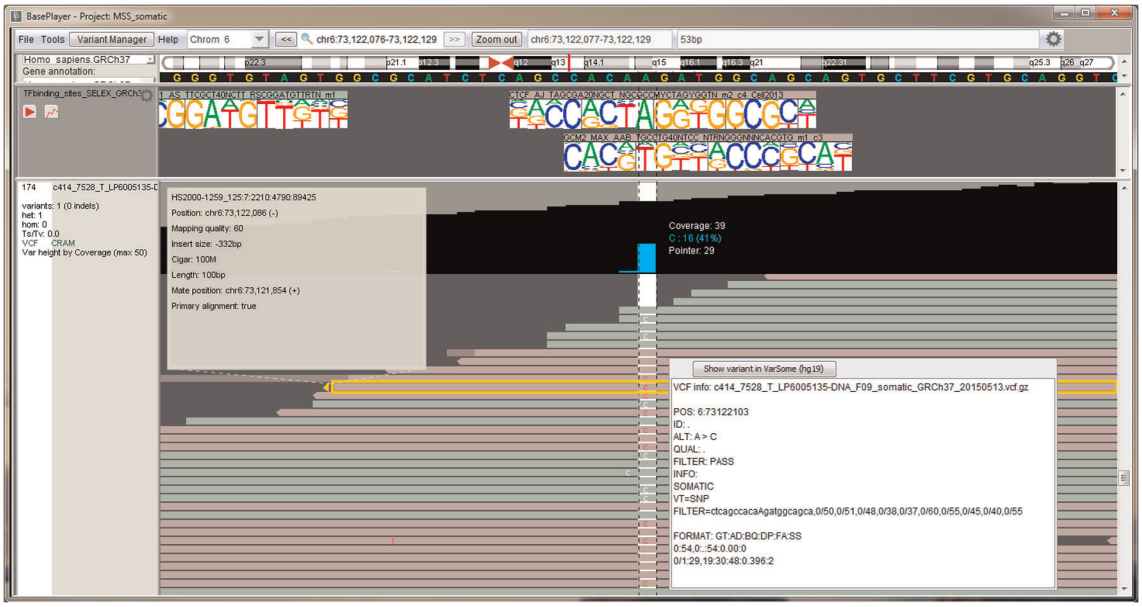

Fig. 8 | Variant quality validation by read-level inspection. Light red and green read colors indicate reverse and forward orientations, respectively. Mismatches in reads are colored accordingly. Read information is shown in a separate dialog (top-left), which is accessed by clicking the read (yellow outline on selected read). Coverage information is located on top of the reads, which shows position-specific coverages and fractions for alternative alleles, derived from the overlapping reads. Read and coverage information is fetched from the BAM file and variant call (white bar behind reads) information from the VCF file. More-detailed variant information is shown in a separate 'VCF info' dialog, if a variant is clicked. For additional annotation, click on the 'Show variant in VarSome (hg19)' button.

Troubleshooting

Troubleshooting advice can be found in Table 1.

Table 1 | Troubleshooting table

Step	Problem	Possible reason	Solution
1	Installer does not start	The installer file may not have the necessary file permissions	Set the execute permission to the installer file and run again. Alternatively, download BasePlayer.zip, unzip it and use Launcher.jar to start the program. The JAR launcher is required for Linux systems. In macOS, click the launcher file while pressing the 'Ctrl' key
2	BasePlayer does not start	Operating system does not have Java Runtime Environment installed	Download and install the Java Runtime Environment from http://www.oracle.com (a 64-bit version is recommended)
	BasePlayer does not start	The launcher file does not have execute permission	Set the execute permission to the launcher file and run again. If the Launcher.jar file is used and does not work, start the program using the command line: java-jar Launcher.jar
	BasePlayer does not start	Installed Java Runtime Environment is incompatible with the requested memory allocation	Download and install a 64-bit version of Java Runtime Environment to allocate >1.2 GB of memory to BasePlayer. We recommend 2 GB for larger projects
	BasePlayer does not start	macOS Gatekeeper does not open applications from 'unidentified developers'	The instructions for dealing with this issue can be found at https://www.imore.com/how-open-apps-unidentified-developers-mac
	Launcher.jar does not launch the software	macOS may have Java compatibility issues	Launch the program with BasePlayer.jar or use the macOS installer and launch with BasePlayer launcher. Note that memory allocation cannot be increased when BasePlayer.jar is used to launch the software

Table continued

Table 1 (continued)

Step	Problem	Possible reason	Solution
3	A genome file cannot be downloaded	Ensembl server is unavailable	Test the server in the web browser by entering 'ftp://ftp.ensembl.org/' in the address bar. If the server is unavailable, try downloading the genome later
	A genome file is corrupted or does not work	Downloading or processing of the genome files may have been interrupted	Delete the genome file by clicking 'File' > 'Genomes' > 'Add new genome', selecting the invalid genome and clicking 'Remove'. Download the genome again. The genome can also be removed manually by deleting the genome folder in the Genome directory
4A(i) and B(i)	VCF file does not open	No read permissions for the VCF or index file	Set the file read permissions on the VCF and the index files
	VCF file does not open	VCF file is not compressed with bgzip or is not sorted	If an uncompressed, unindexed VCF file is opened, BasePlayer will try to create an index before opening the file. An index file will not be created if the VCF file is unsorted or there are no writing permissions for the index file in the folder. Sort the VCF file with the Unix command 'sort -k1,1 V -k2,2n sample.vcf >sample_sorted.vcf'. bgzip file with the Unix command 'bgzip sample_sorted.vcf'. Give the VCF-containing folder writing permissions and open the file again
	VCF file does not open	VCF and/or index file is corrupted	Check the VCF file manually (in Linux, use zless or tabix commands). If necessary, regenerate the VCF and/or index file
	Variants do not show on the sample track	Default filtering thresholds may be too high	Go to Variant Manager and set 'Min. coverage' and 'Min. allelic fraction' to 0. If no variants show up despite this, check that there are variants in the selected chromosome and that the VCF/index files are not corrupt
	BAM file is not recognized	File naming may be incorrect	Make sure that the BAM, VCF and index files share a file name prefix, e.g., 'sample.bam', 'sample.bam.bai', 'sample.vcf.gz' or 'sample.vcf.gz.tbi', and rename the files if necessary
	BAM file does not open, or reads are not shown	Problem may be the same as with VCF files	See the VCF file troubleshooting steps reported above. If necessary, check the BAM file manually using the 'samtools' command
	BED file does not open, or regions are not shown	Problem may be the same as with VCF files	See VCF file troubleshooting steps reported above. If necessary, check the BED file using any text editor
4A(i)	Somatic variants are not shown when opening VCF containing somatic variants	Somatic variants are not found in the last column of the VCF file	Somatic variants should be reported in the last column of the VCF file. Remove all other columns, except for the one containing the somatic variant genotypes
4B(x)	Cannot write an output file	No writing permissions to the folder	Set writing permissions to the target folder
4B(x)	Output file is empty	No variants passed the filters and analysis parameters	If no variants show up in Variant Manager, loosen the analysis criteria and try to annotate the variants again
Supplementary Tutorial 3 (Step v)	Annotation fails when using a CADD track file. Error message: 'Memory allocation exceeded in Annotating variants with <filename>'	Memory consumption exceeds the maximum allocated memory	Increase the allocated memory for BasePlayer. We recommend using at least 2 GB when using a CADD annotation. Alternatively, go to 'Tools' > 'Settings' > 'General', change 'Processing window size (bp)' to 100,000 and press the 'Enter' key
Supplementary Tutorial 3 (Step v)	Cannot right-click or scroll the screen vertically with the right mouse button	The mouse has only one button	In some Mac systems, the mouse does not have a right button. Press the 'Ctrl' key while clicking with the mouse. More instructions are available at https://www.wikihow.com/Right-Click-on-a-Mac

Timing

Step 1, software download and installation: 1 min
Step 2, starting BasePlayer: 10 s
Step 3, downloading and unpacking the human genome FASTA and gene annotation GFF3 files: the duration of downloading these files (~950 MB) depends on the Internet speed (~3 min); unpacking takes 1 min

Step 4, genome-wide variant analysis and gene annotation for ten exome samples (~3 million variants) + gnomAD exome control: 2 min
Step 4, genome-wide variant analysis and gene annotation for ten WGS samples (~50 million variants) + gnomAD genome control: 15 min
Step 4, genome-wide variant analysis and gene annotation for ten WGS samples (~50 million variants) + gnomAD genome control and binding affinity change calculation: 90 min

Anticipated results

Case 1 (Step 4A)

Parameters and sample files used as an example in Case 1 produced a list of five shared variants in five genes. By contrast, variant annotation using the same samples, but without any filtering or other settings applied, would output a list of 9,913 mutated genes. The final variants are very rare in the population, reside in linkage-compatible regions and have been predicted as damaging by M-CAP. In general, the anticipated result in a familial study with at least two cases and population-specific controls is a small set of candidate mutations or genes. These mutations or genes can then be taken forward to functional validation or genetic validation in extended sample sets, for example.

Case 2 (Step 4B)

The analysis conducted in Case 2 resulted in a total of 300 mutation clusters within the ENCODE regulatory regions genome wide, each cluster containing four or more somatic variants. The anticipated result is the detection of recurrently mutated regulatory regions, with mutations potentially affecting the expression of nearby genes. The analysis performed here revealed an unexpectedly high mutation frequency at the binding sites of CTCF/cohesin[11]. In general, BasePlayer can be used in this fashion to integrate genomic information from multiple sources with the goal of narrowing down the set of candidate variants, genes or genomic regions by using TF-binding and/or tissue-specific open chromatin and super enhancer regions, for instance. Indeed, the possibility of using tissue-specific annotation and filtering tracks is an instrumental feature in noncoding variant analysis. Upon visual inspection, the locus shown in Fig. 8 is an example of a robust variant call; both read orientations call the variant, the underlying reference sequence does not contain microsatellite or similar repeats, read mapping qualities are at maximum (60) and reads do not show an excessive amount of mismatches or soft clips.

Inheritance patterns

Inheritance pattern analyses can be performed when at least one individual in the opened samples is set as affected. The following possible inheritance patterns can be returned as results, which assume that no additional annotations or population controls are used: autosomal dominant—heterozygous variants that are shared only by affected individuals; recessive—variants compatible with compound heterozygous, homozygous and X-linked recessive inheritance patterns; and homozygous recessive—homozygous variants shared by all affected individuals such that unaffected individuals must be heterozygous or homozygous reference; compound heterozygous—heterozygous variant combinations in a gene or intergenic regions such that the combinations are not present in unaffected individuals; X-linked recessive—homozygous variants shared by all affected individuals in the X chromosome such that an unaffected father is homozygous for the reference allele; de novo—heterozygous variants found only in the child (both parents must be specified for the child).

In addition, when utilizing the annotation and population controls described in the Procedure, the following results can be anticipated:

Autosomal dominant—with annotation—coding region

Missense/nonsense, heterozygous variants, which are very rare (allele frequency <0.001) in the population and shared only by the affected individuals. These variants are predicted to be damaging to the protein product, are conserved among primates and are recognized by the Cancer Gene Census, as well as being called in regions with good mappability.

Autosomal dominant—with annotation—noncoding regions

Missense/nonsense, heterozygous variants, which are very rare (allele frequency <0.001) in the population and shared only by the affected individuals. These variants are predicted to decrease the affinity score of overlapping TF-binding sites, are conserved among primates and occur in enhancer

regions identified in six cell lines in the ENCODE Project as well as being in DNA regions accessible/open in the epithelial tissue.

Output files

Variant and sequence data can be analyzed and results outputted in various ways, depending on the type of analysis. BasePlayer is able to output a list of variants that survive quality, control and region filtering, and satisfy other criteria such as the requirement for a minimum number of shared variants. The resulting variant lists can be written in TSV-, VCF- or OncodriveFML-specific format. For each variant, the TSV file reports variant annotations such as chromosomal position, base and possible amino acid change, the gene name or flanking genes, quality scores, genotypes, sequencing coverage, allele frequencies and odds ratios (when a control file is used) and annotation track items that overlap the variant (when an annotation track is used). The VCF output enables the user to create filtered VCFs for downstream processing in other tools or, for instance, to create a multi-sample VCF to be used as a BasePlayer control file. BasePlayer also provides a TSV output format accepted by OncodriveFML, which is a tool used to identify cancer driver mutations[31].

Variant statistics can be written to a file when the 'Stats' tab is selected in Variant Manager (Fig. 3a,iv). These statistics include variant counts for different mutation types, the transition/transversion ratio, average allelic fractions and the heterozygous/homozygous ratio. The user can output variant context counts to a separate file for mutation signature analysis by selecting 'only stats' and 'output contexts' in the Variant Manager options (Fig. 3c). Variant contexts are reported as sequence triplets (e.g., TpCpA>TpGpA), followed by the number of respective context-specific variants for each sample[23].

Sequencing coverage statistics for BAM files can be calculated with 'Coverage calculator', accessible in the 'Tools' menu (Fig. 3a,i). The regions to be included in the calculations must be specified using a BED file (e.g., exome target regions). The coverage calculation gives values for average sequencing coverage and mapping quality, as well as the percentage of the total targeted area that is covered. The coverage calculation also provides additional read statistics, such as the proportions of soft clipped and zero mapping quality reads out of all reads.

Reporting Summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

References

1. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The next-generation sequencing revolution and its impact on genomics. *Cell* **155**, 27–38 (2013).
2. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.* **14**, 681–691 (2013).
3. Sabarinathan, R. et al. The whole-genome panorama of cancer drivers. Preprint at https://www.biorxiv.org/content/early/2017/09/20/190330 (2017).
4. Steensma, D. P. et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* **126**, 9–16 (2015).
5. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
6. Alioto, T. S. et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* **6**, 10001 (2015).
7. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
8. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
9. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
10. Donner, I. et al. Candidate susceptibility variants for esophageal squamous cell carcinoma. *Genes Chromosomes Cancer* **56**, 453–459 (2017).
11. Katainen, R. et al. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47**, 818–821 (2015).
12. Kondelin, J. et al. Comprehensive evaluation of protein coding mononucleotide microsatellites in microsatellite-unstable colorectal cancer. *Cancer Res.* **77**, 4078–4088 (2017).
13. Hänninen, U. A. et al. Exome-wide somatic mutation characterization of small bowel adenocarcinoma. *PLoS Genet.* **14.3**, e1007200 (2018).

14. Pradhan, B. et al. Detection of subclonal L1 transductions in colorectal cancer by long-distance inverse-PCR and Nanopore sequencing. *Sci. Rep.* **7**, 14521 (2017).
15. Aavikko, M. et al. Loss of SUFU function in familial multiple meningioma. *Am. J. Hum. Genet.* **91**, 520–526 (2012).
16. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
17. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
18. Jolma, A. et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).
19. Alston, C. L., Rocha, M. C., Lax, N. Z., Turnbull, D. M. & Taylor, R. W. The genetics and pathology of mitochondrial disease. *J. Pathol.* **241**, 236–250 (2017).
20. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
21. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
22. Danecek, P. et al. The variant call format and VCF tools. *Bioinformatics* **27**, 2156–2158 (2011).
23. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
24. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
25. Milne, I. et al. Tablet—next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2009).
26. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**, 464–469 (2011).
27. Fiume, M. et al. Savant Genome Browser 2: visualization and analysis for population-scale genomics. *Nucleic Acids Res.* **40**, W615–W621 (2012).
28. Abeel, T., Van Parys, T., Saeys, Y., Galagan, J. & Van de Peer, Y. GenomeView: a next-generation genome browser. *Nucleic Acids Res.* **40**, e12 (2011).
29. Wöste, M. & Dugas, M. VIPER: a web application for rapid expert review of variant calls. *Bioinformatics* **34**, 1928-1929 (2018).
30. Kallio, M. A. et al. Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics* **12**, 1 (2011).
31. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).
32. Smedley, D. et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.* **10**, 2004–2015 (2015).
33. Fu, Y. et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15**, 480 (2014).
34. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
35. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2014).
36. Jagadeesh, K. A. et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581–1586 (2016).
37. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
38. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–719 (2011).
39. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
40. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
41. Hoffman, M. M. et al. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41**, 827–841 (2013).
42. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44**, D110–D115 (2016).
43. Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181–3182 (2009).

Acknowledgements

We thank T. Kivioja for his guidance in regard to the SELEX data and A. Ollikainen for the voice-over in the demonstration videos. We thank B. Pradhan and L. Kauppi for sharing their unpublished Nanopore data. We also thank M. Aavikko, L. van den Berg, D. Berta, O. Kilpivaara, J. Kondelin, H. Kuisma, Y. Li, M. Mehine, H. Metsola, J. Ravantti, L. Sipilä, T. Tanskanen, P. Vahteristo and N. Välimäki for testing BasePlayer and giving suggestions and additional support. We acknowledge ZeroTurnaround for creating the JRebel plugin for Eclipse (IDE). This work was supported by grants from the Biomedicum Helsinki Foundation; the Cancer Society of Finland; the Emil Aaltonen Foundation; the Juhani Aho Foundation for Medical Research; the Sigrid Juselius Foundation; the Academy of Finland (Finnish Center of Excellence Program 2012–2017, 250345); the European Research Council (ERC, 268648); a European Union Framework Programme 7 Collaborative Project (SYSCOL, 258236); the Nordic Information for Action eScience Center (NIASC); and a Nordic Center of Excellence grant financed by NordForsk (62721 to K.P.).

Author contributions

R.K. designed and developed the protocol. R.K. and E.P. wrote the protocol. I.D. contributed to writing the protocol. I.D., T.C., E.K. and K.P. assisted in developing and testing the software. E.P., V.M. and L.A.A. supervised the research.

Data availability

No previously unpublished data sets were generated or analyzed during the current study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41596-018-0052-3.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.K. or E.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published online: 15 October 2018

Related links

Key references using this protocol

Katainen, R. et al. *Nat. Genet.* **47**, 818–821 (2015): https://doi.org/10.1038/ng.3335
Pradhan, B. et al. *Sci. Rep.* **7**, 14521 (2017): https://doi.org/10.1038/s41598-017-15076-3
Donner, I. et al. *Genes Chromosomes Cancer* **56**, 453–459 (2017): https://doi.org/10.1002/gcc.22448